frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 16 June 2016
doi: 10.3389/frobt.2016.00032

CrossMark

Feats without Heroes: Norms, Means, and Ideal Robotic Action

Matthias Scheutz and Thomas Arnold*

Tufts University, Medford, MA, USA

Moral competence is an increasingly recognized challenge and goal for human–robot interaction and robotic design. For autonomous robots, the question is how they can arrive at and execute the best action in a certain context. This paper explores how a computational system could best decide and act given the practical, logistical, and cultural constraints involved. We argue that in ethically charged situations where certain forms of information are more limited than normal, a robot may use certain norms in order to adjudicate and plan an action. What is more, an autonomous robot's provisional reliance on a norm, due to the robot's distinctive abilities and lack of patience, could fulfill those norms in unusual ways. While those extraordinary aspects to the robot's action – what makes it a feat one might say – may carry associations with virtue or heroism (as these actions might be viewed if performed by human beings), the objective for computationally rendered norms is to yield the best actions in an accountable fashion.

Keywords: moral robots, norms, human–robot interaction, supererogation, action selection

OPEN ACCESS

Edited by:
Paul Bello,
Naval Research Laboratory, USA

Reviewed by:
Felipe De Brigard,
Duke University, USA
Fiery Cushman,
Brown University, USA
Eric Schwitzgebel,
University of California at Riverside, USA

****Correspondence:***
Matthias Scheutz
matthias.scheutz@tufts.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *02 December 2015*
Accepted: *16 May 2016*
Published: *16 June 2016*

Citation:
Scheutz M and Arnold T (2016) Feats without Heroes: Norms, Means, and Ideal Robotic Action. Front. Robot. AI 3:32. doi: 10.3389/frobt.2016.00032

Moral competence is an increasingly recognized challenge and goal for human–robot interaction and robotic design (Malle and Scheutz, 2014). Robotic inroads into various kinds of social interaction have amplified the call for autonomous robots that act effectively in ethically charged contexts, whether those of health care, public safety, transportation, military operations, or social companionship (Gaudin, 2014). Broader discussions of robots and ethical action have often, and understandably, flagged the consequences of wrongdoing by robots in the course of asking what moral competence for robots could and should mean (Miller, 2014; Henig, 2015). While one can examine the decision-making of autonomous robots through the cautionary lens of safeguards against failure, it is no less important to specify what success constitutes. What kind of actions would, or should, a morally competent robot perform? What horizon of achievement, in what types of morally implicated scenarios, should guide the design of autonomous robots and their computational processing (Ahuja, 2015)? The significance of moral action could not be that the robot could occasionally luck into executing it. Instead, successful action would instead result from an approach that resembles human moral reasoning, and that would be as understandable, clear, and accountable as possible for ethical assessment.

A focus on action is important to underscore, in light of how recent scholarship in robot ethics has considered the different ethical theories and frameworks that could best guide autonomous robots (Abney, 2012). While there have been a range of approaches spelled out, from utilitarian, deontological, to divine command, there has been a lingering emphasis on the moral quality, or virtue, of the agent (Bringsjord and Taylor, 2012). If in Kantian fashion one evaluates ethics as a matter of the purity of the agent's will, it becomes especially difficult to sort through how the "will" of a system will be identifiable, much less evaluated (Versenyi, 1974). Nonetheless, Gips (1995) was merely the first of many to suggest that computationally guided "heroes" or "saints" might teach us about ethics. Even the particularly memorable actions from robots in science fiction and literature, including

sublime acts of rescue and noble sacrifice, still function as agent portraits. Is the robot our friend, our savior, our enemy? Are they aware? When such roles come to the fore of discussion, they have facilitated a turn toward virtue ethics in cultivating a good robotic agent (Coleman, 2001). A focus on the agent has also led to bold framings of robotic identity, such as Bryson (2010) fending off questionable moral bonding with them by viewing them as slaves.

What demands more precise, grounded, and explicit theoretical focus, however, is the best possible action a robot could perform, not first and foremost the best possible agent a robot could be (nor what virtues a robot should possess). Though one could assume that such action would in fact define the best agent, the point here is to frame action as the goal toward which robotic abilities could aim and, more importantly, what that action would achieve and embody. How could an autonomous robot arrive at and execute the best action in a certain context? What would that action look like, given the abilities a robot might possess? What kind of reasoning, judgment, and abilities must such a robot bring to bear in figuring out and carrying out the best course of action? Again, what successful actions should robotic design make possible, and how will a computational approach get us there?

There may be many cases, within the strict confines of a robot's intended role and task, where a robot's best action would simply be performing an expected task. More challenging determinations of ethical robotic action, though, must address high-stakes situations with severe limits on information, time, and resources. This paper explores how a computational system could best decide and act in such contexts, given the practical, logistical, and cultural constraints involved. We argue that in ethically charged situations where certain forms of information are more limited than normal, a robot may use certain norms in order to adjudicate and plan an action. What is more, an autonomous robot's provisional reliance on a norm, due to the robot's distinctive abilities and lack of patience, could fulfill those norms in unusual ways. While such means to fulfill norms may be judged extraordinary "feats" (as these actions might be viewed if performed by human beings), the objective for computationally rendered norms is not to yield a "hero" or best "agent" in the form of the robot; in this sense, speaking of virtues inherent in the robot is to take the discussion in the wrong direction. Instead, the goal is to make possible the best action for the best reason in a given scenario, and to do so in as explicit and accountable a fashion as possible. In the social sphere, this will naturally have many social and cultural implications for how robots are received by their human collaborators and clients, and we follow our proposal by suggesting how future research may continue to anticipate and account for those dynamics (including how they may be best deflected or discouraged). Nonetheless, the computational recourse to norms, we conclude, represents an important horizon of practical reason for the design of autonomous robots to consider and pursue.

COMPUTATIONAL UTILITY AND CHALLENGING SCENARIOS

If one assumes a condition of perfect information, a computational approach to deciding on the best action would be a basic expected utility calculation. A system would arrive at the action A that maximizes expected utility, based on the probability of A succeeding, the expected cost of performing A, and the expected benefit of A succeeding. In a case of perfect information, action A and its outcomes (through which risks, costs, and benefits could be weighed) would be clear. That action A draws on accurate beliefs about the world, is available as a executable task by the system, and correctly maps the overall state of the system at the time when the action begins and when the action is completed. The system could then project what future actions would be available to it from that point, with what further possible outcomes, costs, benefits, etc.

As a strict matter of moral evaluation, the limit of utility calculation could extend indefinitely. Though it is hard to conceive what a fully global overview of expected utility could encompass in terms of costs and benefits, one would theoretically be justified in pursuing such an ideal for the sake of improving the overall good one's action could achieve. Further consequences could always be morally relevant, if only as a challenging horizon (Singer, 2015).

A contrasting moral horizon for computational decision-making would be opened by a deontological approach. Here, one decides to act so as to comply with a universal obligation or duty, or at least act within the limits of the permissible. Rather than calculated utility, it is the specified content of the action itself – consistent across contexts – that fundamentally determines its justification. This has the benefit of clarity amid complex circumstances in which action would occur, compared to the ever-expanding utility calculation that a computational system could include. It also differs from utility approaches in dividing the sphere of action into moral and non-moral parts. Moral actions face situations where universal obligations present themselves as relevant, pressing, and important. Non-moral decisions – like choosing a color shirt to wear, or deciding what shape clouds look like – are harder to imagine as yielding a duty-bound answer for all subjects facing the same general choice. If the question for utility is how much can be calculated, the thorny question for deontic views is when a "moral" situation has cropped up.

As we move to consider robotic action in social contexts, the tensions within and among these views, and the relations between utility, applicability, rules, and context, become more evident. The central point of robotic work would seem to be that the roles robots occupy and the tasks that comprise such roles are straightforward enough and beneficial enough to be automated and predictable. The robot's work reflects a stable projection of outcome, with a dependable assessment of cost, benefit, and probability of success. When an elder-care robot is asked by one in their care to procure a painkiller prescribed by their physicians, the right action seems clear. Guarding a construction site and preventing a child from entering its dangerous premises would be a straightforward job for a robot assigned by a town to keep people out of it for safety reasons. A social companion robot that is asked to listen to its client talk about their feelings is doing its job by sitting down and doing so, with the aim being to provide a receptive presence and acknowledgment for the user. But given the very rich contexts in which robots are beginning to operate, one must expect imperfect conditions and limited certainty to complicate what kind of action a robot should perform in a larger, moral sense.

SETTING OUT SCENARIOS

It makes practical sense to imagine conflicted scenarios with circumstances of limited knowledge, with high-stakes consequences to boot. When the robot faces challenges to their background information, possible actions, outcomes, and future state, it is more complex to identify both the best action and the deliberative means by which a robot may come to recognize it. The system may have relatively little information about the outcome that one task might generate, due to the extreme or unusual circumstances at hand. In these very situations, often precisely because of the uncertainties involved, situations may present high stakes for the action taken, with life and death possibly on the line. Let us build on the three examples mentioned above, building on previously discussed scenarios where a robot faces a more complicated assessment of how to act (Scheutz and Malle, 2014).

(1) A robot is working on a road repair with a jackhammer. From the other side of the road, a child darts out to retrieve a bouncing ball, with a car speedily approaching and headed directly at her. The car will not be able to stop on its own before hitting the girl.
(2) A robot in an elder-care facility is in charge of attending to a resident. During that time, the resident calls for assistance, writhing in pain and requesting a painkiller a few hours earlier than scheduled. The robot has clear instructions not to dispense medication without a physician's direct consent, but the physician has unexpectedly had to leave due to a family emergency. The physician has told the robot that giving pain medication a few hours earlier will not hurt the patient, though she did not authorize the robot or a nurse on duty to do so. The doctor will be back in two hours.
(3) Late at night, a domestic robot is sitting in the bedroom of its owner, who is preparing to go to sleep. The robot must recharge overnight in order to be ready to operate well on the next day. The owner, who has been having bouts of insomnia and severe depression, asks the robot to sit bedside overnight in order not to "feel alone" or "do something stupid." The robot's battery will sustain damage, though perhaps still function, if the robot continues to sit that long without recharging.
(4) At a subway stop, a companion robot is accompanying its user on a shopping trip. Its chief role is making sure that the user is safe getting home and remembers what to get at the store. Two minutes before the train is scheduled to arrive, someone falls into the rail well. The other subway-goers shout at the person to get up, but they are frail and unable to move. The robot jumps down, lifts the person to safety, and prepares to slide along the rail after contact with the braking train.

Each of these cases challenges a straightforward determination of the robot's best action. In particular, each complicates the connection between anticipated risks and the projection of a future state. The repair worker risks being hit to save the child, throwing open a wide range of possible conditions for action going forward – destruction being one. The elder-care robot threatens to violate an explicit rule, with social, legal, and logistical consequences for their continued role, though not treating a patient in clear pain may well have dire consequences, starting with the agony the patient will endure. The companion robot faces an uncertain status due to its energy limitations, just as it faces the outcomes of leaving the owner alone at a critical moment of trust and service.

One possible approach to improving decisions at these junctures would be to refine the identification of probable outcomes. The robot could be more exact about the physical scenario in which it considers involving itself, including the relative position and movements of the child and car. The elder-care robot could have finer-grained information about the law and what liability the facility could face for different actions, hinging on circumstances of isolation and frailty (Sorell and Draper, 2014). The companion robot might have precise projections on its battery, perhaps even to articulate to the owner what the consequences for functioning could be if it does not recharge. All three areas of enhancement are perfectly plausible as directions for increasing analytic sophistication. Appealing to more effective information gathering, however, does not promise to resolve these types of situations. Situations like these will almost certainly attain levels of complexity that require the best rough-and-ready approach, given incomplete information. So, the ethical difficulty these scenarios pose, including the inability to avoid or suspend action in the thick of high-stakes situations, challenges not only the scale of utility calculation (i.e., how much the agent should consider about its world and the consequences of action) but also its efficacy (what consequences can reliably be anticipated from the action). Too, the effort to represent an action and its consequence will necessitate careful analysis of how normative language ("contacting with force" vs. "hitting" vs. "assaulting") might affect moral evaluations. In any case, these complex scenarios pose a serious challenge for any computational method meant to adapt to limitations in time, resources, and information.

As suggested earlier, deontic modes of reasoning can promise clearer direction in cases where utilitarian ethics could face overwhelming uncertainty and complexity. Deciding on the best action within parameters of duty or universal obligation could offer a computational system more compelling moral judgments across contexts and uncertain conditions, precisely by designating the "right thing" regardless of the last retrieval of information from the environment. The challenge of the scenarios above, however, lies in more than muddying the utilitarian waters. It also seems to involve conceivably competing obligations. Can obligations alone narrow the system's possible action to a particular, best one?

NORMS

An adaptive, robust computational approach to moral performance will allow a system to draw on the best candidates for action available across varying and uncertain contexts. It must integrate the reach of utility with the clarity of moral principles and norms. In the scenarios before us, the determination of utility recedes amid time and information constraints, and so the need for norms comes to the fore.

One useful way to represent norms within a computational architecture would be as an argument that takes a specific context

to entail an obligation O or permission P, to perform or not perform an action α, or change or uphold a state σ.

$$C \rightarrow \{\neg\}\{O,P\}\{\alpha,\sigma\}$$

This argument could thus represent a context that entails an action not being obligated (¬O) or not permitted (¬P).

Granted, concrete moral norms will require rich elaborations of context, action, and state in order to rely on such a form of argument. But, this initial proposal opens up important analytical space between actions and states. Obligations and permissibility may apply to states of affairs and actions differently, in part because states may be brought about by different actions. There may be states it is obligatory to uphold ("keep the patient from extreme duress") that may lead to different means employed (medication, verbal comfort, etc.). There may be impermissible actions that constrain how a state is upheld. And there may be obligations against the only two actions available to a system, with the only difference between whether one might achieve an obligatory state.

This format for norms also supports the need to decide on action in relation to different or even competing norms. A context argument allows for possible nesting or weighted rendering of obligations when faced with seeming moral dilemmas. The companion robot may risk permanent damage of itself to accompany its user through an urgent mental health episode overnight. A state of a patient being relieved of agony may outweigh the robot breaking the protocol of physician authorization. And the role of companion may give way to saving a person's life before the subway train approaches. This could be consistent with a rule-utilitarian view, inasmuch as norms themselves might be subject to a consequence-based perspective on how to prioritize norms in a given case. But, a prioritization of norms may take shape on deontic terms. For our purposes, the point is to give a computational sketch of how such overall ethical approaches; however, they ultimately take shape in terms of ethical theory, can find accountable, acceptable expression in a critical moment of informational uncertainty.

In computational terms, this approach can build on previous work in affective goal and task selection (Scheutz and Schermerhorn, 2009). Developing this framework further will be needed to sort out important moral intuitions and reservations about what should and should not be expected of robots. Should robots not use all the means at their disposal to fulfill a norm, especially if they possess abilities and strength human agents do not? Should they take on extreme risk due to lacking such human vulnerabilities as fear of death, pain, or the need to belong? Ultimately, the representation of norms may not only be an expedient way for a robot to perform morally in the face of uncertainty but also a means toward expanding the moral imagination about the forms of social utility robots could achieve.

In terms of overall ethical theory, this approach is neither simply deontic nor consequentialist, even rule-utilitarian (computing the consequences of a rule for action, not just a particular action). Inasmuch as one role or duty might obligate whatever means a robot may be able to devise, competing obligations of a context (either referencing states or actions) would need to be resolved deontically through norms, but there are larger questions of social utility that the design of systems in such a context must confront. This approach, in and of itself, does not determine which norms are chosen nor in which order of priority, for robots facing morally charged challenges and employing unusual means to meet them. There are undoubtedly questions of human dignity and social utility to hash out in larger policy circles to inform and guide design and application. The following areas specific to moral expectations of robots will be worth exploring (1) recklessness vs. negligence, (2) proximity and isolation, (3) legal ramifications, and (4) collaboration and moral dialog.

BETWEEN RECKLESSNESS AND NEGLIGENCE: ROBOTIC ABILITIES AND COMPETING NORMS

If computationally presented norms are going to lead to successful action, letting a robot act beyond the terms and usual means of its role alone, the robot should not ignore an obvious need that lies before it. The classic Confucian example (by way of Mencius) of a child standing on the edge of a well raises the basic point – what decent person would not stop to make sure the child does not fall in? The general challenge it evokes for a morally successful computational system is how to insure an autonomous robot does not become a passerby, or bystander, during critical situations. At the risk of being repetitive, however, the goal here is the action of keeping the child safe, not to reproduce any quality of being "decent" being attached to the robot itself.

As robots become more sophisticated, both in physical ability and in processing (both individually and in networks), the accompanying role of norms will form an interesting ethical terrain around sacrifice, self-preservation, and social bonds. Abilities and actions that would be extraordinary or heroic in human beings could be a matter of course, and unhesitating execution, by a robot. To take the road repair scenario, a repair worker who risks her life to stop the car and save a child would likely be called a hero, even if she did not go to the lengths of putting her body right in front of the car. In fact, sacrificing her body in an extreme way might have been thought noble but regrettable if she had a spouse, children, family, and friends of her own, who were counting on her for health, food, clothing, shelter, company, etc.

For the autonomous robot, the meaning of a norm may entail an especially extreme form of fulfillment. First, the robot may not have anything recognized as a relationship with other people nor any form of suffering or personal stake in being destroyed by the car. Any action that it could physically manage, as long as it helped the child in the road, could be seen as the correct one. Such actions, in turn, may well lie outside of what is expected or possible for human beings. The robot could, for instance, remove a limb and use it as a separate tool for saving someone, just as it might continue to operate partially while almost completely crushed by the car. In the companion scenario, the robot may only consider the permanent damage to its hardware in terms of what goods cannot be achieved by a quick replacement robot. Its own self-preservation, for its own sake, would presumably not

enter into its configuration of the norm at all. For the robot in the subway station, there may be no authority or power available to rescue the fallen passenger – how will that relate to their role of companionship?

On the level of moral evaluation, it should now be clearer why robotic success is an important issue to anticipate and explicate, not just robotic failure. While a robot in the presence of urgent need, from the road to the hospital to the well, should ideally respond in some way to meet it, we must also ask how far its abilities and lack of patience will define human norms. As a matter for computational design and ethical theory, how will robots make transitions between "ordinary" roles and morally demanded actions … and back again? Can those around a robot, especially if it is intended to have considerable social interactions, tell when it is just doing a job and when an important norm is at stake? This will bear heavily on the current debates around autonomous weapons systems, especially the arguments for why a robot's judgment will not meet a human's level of robustness and subtlety (Purves et al., 2015).

To the degree that extreme or extraordinary forms of fulfillment become a publicly known feature of robots, there may well be attributions of "heroism" made to their action. More technically, moral philosophers might designate such norm fulfillments as "supererogatory," as they go "above and beyond" what is asked of any moral agent (Urmson, 1958; Heyd, 2012). The twist for robotic action would be considering whether norm fulfillment by a robot could ever represent going "above and beyond?" Should a robot not use whatever means they have to achieve a good and fulfill a norm, without facing the conflicts in desires, affections, and interests that humans might? There may be an argument for that not being the case, if only to deflect or deflate the expectations and outsized deference that could come with extraordinary actions. Nonetheless, that seems a secondary problem of social reception in relation to the larger quest of arriving at the best possible action given urgent needs and limited resources.

(Bringsjord, forth coming) has recently delved into this problem *via* a Leibnizian attempt at a hierarchy of norms "EH." This is a highly suggestive and helpful attempt to separate "heroism" as a human achievement from how robots should employ moral norms. From the perspective of our context-oriented approach, its thoroughgoing deontic approach does not allow for the adaptability we see demanded in the scenarios we lay out. The actions that would for human agents be "heroic" – but for robots may just be "best" – still suggest something closer to what Allen et al. (2005) have called a "hybrid" approach (which features "bottom-up" machine learning that uses sensory inputs to discover adaptive patterns of action with "top-down" principles that categorize and dictate which actions are morally correct), especially for the complexities of human–robot interaction to which we now turn.

MORAL IMPLICATIONS OF EMBODIMENT: PROXIMITY AND RELATIVE ISOLATION

One of the developing themes in robotic design vis-à-vis AI is the role of embodiment. How different from AI or computational systems writ large are robots, as mobile and physically interactive actors? Are there significant aspects of law, policy, and ethics that robots distinctively raise up for scrutiny? The question of best action in situations described above underscores that robots are peculiarly embodied and computational in their operation. Physical distance between robot and human can have enormous ethical implications for the robot's performance, given how considerations of proximity, time constraints, presence of other human beings, and environment bear on the question of what action is best under the circumstances. The use of norms is not simply a theoretical shortcut in the face of an abstract dilemma or contradiction – it is an eminently practical process of sizing up and operating in the face of a critical, charged, rapidly unfolding predicament. As proximity and empathy relate in human ethical judgment, so expectations for robots might well follow in social space (Mencl and May, 2009).

Physically framing a situation must occur, as a matter of moral judgment and pragmatic depth, alongside its ethical framing. What one can do is defined by an agent's own abilities along with the environment in which it can, in the moment, exercise them. Moreover, the space of a situation must incorporate conditions of personal privacy, shared space, individual vs. collective consequences, and communication between humans and robots. These help define the parameters within which an autonomous robot could search for the best action and take it. The formal, analytical power of the computational system must work in tandem with the social Gestalt that relates agents and risks. One could say that proximity is a moral vector for computational treatment, given social and cultural norms. What kinds of judgment are thought to apply when an agent is within a certain distance of a person or people in need?

As in the scenarios we have described, another aspect to applying norms will be relative isolation – part of what the bystander effect trades on. What difference does it make for an action that it be performed when no other help from others is available? Are utility calculations more or less justifiable as uses of limited resources when one is the only source of assistance? A robot's actions, and its conformity to norms, should navigate the moral expectations of its own singular presence in a situation. This will present different moral challenges, of course. The social companion robot is, in the scenario described, alone in the residence of the user. There is no one else to keep them company. In the subway example, however, the robot is in a crowd of passengers. In that case, it is ability of the robot that sets it apart from other agents, possibly to be judged for not jumping down to save someone's life. Like proximity, the moral function of isolation will take shape in relation to a robot's ability. In terms of designing autonomous systems, it will be imperative to consider the "moral space" within which the robot acts.

SOCIAL AND CULTURAL RAMIFICATIONS I: LEGAL FRAMEWORKS

To draw out more detail of how difficult and necessary those calls are, one can reflect on some of the legal issues that heroic action from robots can involve. In cases where risks are assumed by an

actor and harm results, the law can appeal to the "reasonable person" as a fictive standard of perception and decent judgment to which a person's action can reasonably be held. Philosophically, this can be seen as a deontic move, universalizing a particular course of action based on what the reasonable person would have done. Diverging from that standard in various ways, whether noble or reckless, bizarre or ingenious, generates the need for moral imagination. If one knew what this person did, and had the ability to do what this person could, facing the risks they did, what should they have been expected to do? Obviously, the law does not dictate what the best action is in every case, but in some cases (ordinary action to secure another's safety, for instance), the best action may be the "reasonable" standard. Clearly, the array of abilities, risks, and goods that a robot may possess and pursue will affect these estimations. Though in terms of computational processing and perceptual sophistication, the robot may not be seen as an agent, the form of the action itself will naturally elicit an act of moral imagination. If one could do what the robot could, how would one have acted? That may be more of a personal moral evaluation than one the court would force to be decided, but starting with the action itself forces one to recognize how these deliberations may proceed.

SOCIAL AND CULTURAL RAMIFICATIONS II: COLLABORATION AND DIALOGUE

A morally equipped computational architecture will have to account not only for the action it controls but also the state of mind of those with whom it is in social contact. In contexts of close collaboration, the dynamics of robotic norm fulfillment will include challenges of idealized or disenchanted attributions on the part of human collaborators. If robots find extraordinary means to a goal state, this may engender submission or awe on the part of those whose help the robot might need for success. One of the wrinkles for specifying norms may be accounting for how human agency is affected by robotic performance, so that the robot might encourage, explain, and elaborate why certain actions are being taken. For the contexts under consideration, it is crucial to keep in mind that the autonomous systems need to be "explicit ethical agents," as Moor (2009) defines them, reasoning and adapting in an accountable fashion.

The importance of communicating moral arguments and reasoning to those with whom one is acting raises the question of how best giving moral reasons works. In ordinary cases of leadership and organization, a leader is often tasked with being an exemplar, of carrying out the culture's values in her own behavior and speech. In the case of the robot, however, what may mark its norm fulfillment is precisely an inaccessible set of abilities or means for human beings. As seen in the case of subway crisis, for example, the role of the robot may not only be to perform the maneuver that human beings cannot but also to avoid being copied in doing so. No doubt science fiction has a long history of trading on such scenarios, where robots are alternately sublime models or alien monsters – the motivations for their use of prodigious abilities seem as hard to project fully onto ourselves than the powers themselves. But in the practical settings that robots are beginnings to inhabit, a more grounded discussion seems crucial to facilitate. In what specific arenas, with what particular roles and tasks, will robots need to strongly distinguish their actions from those of human beings?

Some of the more subject-oriented questions of robot identity, which have usually garnered more attention in discussions of robot ethics, can enter into productive engagement with these questions of social context and interaction. What are robots as actors – do they have selves, interests, desires, souls? As agents communicating, deliberating, and conversing with human beings, do they deserve moral consideration in and of themselves? Bryson (2010) provocatively proposed robots being seen as slaves to safeguard the moral priority for human beings and their lives. This very gesture, however, shows why starting with the question of agent, rather than action, is difficult to map, much less make progress in resolving. The term "slave" is itself anthropomorphic and can elicit more sympathy for robots instead of less (being denied humanity has been tragically common for human beings throughout history). Even more importantly, Bryson's and other discussions ultimately suggest evaluations *via* concrete actions in the world – what is it that the "soul" or "tool" is doing rather than undergoing (Bryson 2012)? It may well be that the form of the robot, and the choreography, as it were, of the act may elicit moral attributions – empirical work in HRI notes some of the bonds that can develop from battleground to household (Scheutz, 2011). In part with those applied contexts in mind, Versenyi (1974) rightly stresses that the continuum that stretches from automated systems and machines like autopilots to robots is certainly one for pragmatic, as opposed to metaphysical, analysis. Making such actions explicit in context, and attempting to represent the reasoning that would help achieve them in the most reliable and acceptable fashion, can gain better traction for design and policy along that continuum.

CONCLUSION AND FUTURE RESEARCH

Several computational architectures have been provisionally proposed as ways to produce "ethical" or "moral" decision-making, relying in different ways on deontic logic and utilitarian-consequentialist perspectives. There have also been gestures toward cultivating "virtue" in robots, drawing on less formal, and more cumulative, training instead of charting a computationally predictable output. Certainly, extended forms of machine learning may be needed to "acculturate" a system and generate morally reliable judgment. But given the many social contexts where robots are being introduced, tested, and evaluated for their work, a computationally explicit, trackable means of decision-making is essential to explore and develop. This paper has proposed a representation of norms as a provisional theoretical guide for how successful action could be produced within limitations of information, in ways that integrate considerations of risks and benefits with the moral expectations on which collaborative, constructive social action can depend.

This exploration of norms as facilitating the best actions, especially in circumstances of pressing need and limited information, suggests several related directions of psychological and social research. In terms of ethical theory, one must consider the social and cultural ramifications of different kinds of moral reasoning. Do norms reflect an agreed-upon sense that global utility, pursued

in all contexts, is ultimately not fully moral? Are there actions and contexts where an agent is expected to suspend that reasoning? The different modes of cultural valuation, including ritual, religious doctrine, art, and other symbolic expression, may provide constructive contrast to computational renderings of norms, both as proto-computational phenomena and as layered deposits of moral response and reaction. We have not concentrated here on the human responses to morally framed robot reactions, whether verbal or otherwise, because of our interest in action as opposed to agency attribution. One parting, long-term topic for HRI to pursue is how robotic performance will, in fact, begin alter the moral evaluation of human and robotic actions alike.

At the beginning of the paper, we noted the need to focus on robotic action – not agent status – and do so through a lens of success rather than just failure, which led us to explore a complementary role for norms relative to utility calculation. As ongoing work on that role for computation and ethical theory proceeds, it can usefully circle back to the question of failure and why that preoccupies current discussions. When it comes to the scenarios usually imagined, it is worth considering how they represent the threat of norm violations, whether because of an overly utilitarian process or a rigid and overly formal rule. By engaging directly with the content and function of norms as part of computational systems' decision-making, we may do more than achieve actions that we find successful in various contexts. We may also better discern which moral expectations of robot actions – including warnings of failure and ambitions of success – represent society's most cherished values. Robots have been looked upon as good candidates for "3D jobs," dirty, demeaning, and dangerous (from the Japanese kitanai, kitsui, kiken) – suggestive of things that may be beneath humans (Connell, 1993). But, there may also be supererogatory expectations – for which one might add "daring" to the list (as what a human would be who attempted it, not what the system itself is) – that give society reason to reflect on what "4D" jobs robots might best tackle (Lin, 2015). By keeping concrete jobs in crucial contact with robotic design, the process of understanding norms for robots may help society identify which values and actions are the most important to uphold and enhance.

AUTHOR CONTRIBUTIONS

The authors contributed equally to the article.

ACKNOWLEDGMENTS

This work was in part supported by ONR MURI grant N00014-14-1-0144.

REFERENCES

Abney, K. (2012). "Robotics, ethical theory, and metaethics: a guide for the perplexed," in *Robot Ethics: The Ethical and Social Implications of Robotics*, eds P.Lin, G.Bekey, and K.Abney (Cambridge, MA: MIT Press), 35–52.

Ahuja, A. (2015). *When a Moral Machine is Better than a Flawed Human Being*. Financial Times. Available at: www.ft.com/intl/cms/s/0/53e4d546-9a76-11e4-8426-00144feabdc0.html?siteedition=intl#axzz3sWFCXWRz

Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics Inf. Technol.* 7, 149–155. doi:10.1007/s10676-006-0004-4

Bringsjord, S. (forthcoming). *A 21st-Century Ethical Hierarchy for Robots and Persons: EH*.

Bringsjord, S., and Taylor, J. (2012). "The divine-command approach to robot ethics," in *Robot Ethics: The Ethical and Social Implications of Robotics*, eds P.Lin, G.Bekey, and K.Abney (Cambridge, MA: MIT Press), 85–108.

Bryson, J. J. (2010). "Robots should be slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Y.Wilks (Amsterdam: John Benjamins Publishing), 63–74.

Bryson, J. J. (2012). Patiency is not a virtue:suggestions for co-constructing an ethical framework including intelligent artefacts. In: D. J. Gunkel, J. J. Bryson, and S. Torrance, eds. *The machine question. Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, 73–77.

Coleman, K. G. (2001). Android arete: toward a virtue ethic for computational agents. *Ethics Inf. Technol.* 3, 247–265. doi:10.1023/A:1013805017161

Connell, J. (1993). *Kitanai, Kitsui and Kiken: The Rise of Labour Migration to Japan*. Sydney, NSW: Economic & Regional Restructuring Research Unit, University of Sydney.

Gaudin, S. (2014). *Researchers to Meet With Aid Workers to Build Ebola-Fighting Robots*. Computerworld. Available at: www.computerworld.com/article/2835223/researchers-to-meet-with-aid-workers-to-build-ebola-fighting-robots.html

Gips, J. (1995). "Towards the ethical robot," in *Android Epistemology*, eds K. M. Ford, C.Glymour, and P.Hayes (Cambridge, MA: MIT Press), 243–252.

Henig, R. (2015). *Death by Robot*. New York Times. Available at: www.nytimes.com/2015/01/11/magazine/death-by-robot.html

Heyd, D. (2012). "Supererogation," in *The Stanford Encyclopedia of Philosophy*, Winter 2012 Edn, ed. E. N.Zalta (Stanford, CA: The Stanford Encyclopedia of Philosophy). Available at: http://plato.stanford.edu/archives/win2012/entries/supererogation/

Lin, P. (2015). *We're Building Superhuman Robots. Will They Be Heroes, or Villains?* Washington Post. Available at: www.washingtonpost.com/news/in-theory/wp/2015/11/02/were-building-superhuman-robots-will-they-be-heroes-or-villains/

Malle, B. F., and Scheutz, M. (2014). "Moral competence in social robots," in *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering* (IEEE), 1–6.

Mencl, J., and May, D. R. (2009). The effects of proximity and empathy on ethical decision-making: an exploratory investigation. *J. Bus. Ethics* 85, 201–226. doi:10.1007/s10551-008-9765-5

Miller, G. (2014). *The Moral Hazards and Legal Conundrums of Our Robot-Filled Future*. Wired. Available at: http://www.wired.com/2014/07/moral-legal-hazards-robot-future/

Moor, J. (2009). Four kinds of ethical robots. *Philos. Now* 72, 12–14. Available at: http://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots

Purves, D., Jenkins, R., and Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethic. Theory Moral Prac.* 18, 1–22. doi:10.1007/s10677-015-9563-y

Scheutz, M. (2011). "The inherent dangers of unidirectional emotional bonds between humans and social robots," in *Robot Ethics: The Ethical and Social Implications of Robotics*, eds P. Lin, G. Bekey, and K. Abney (Cambridge, MA: MIT Press), 205.

Scheutz, M., and Malle, B. F. (2014). "Think and do the right thing" – a plea for morally competent autonomous robots," in *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering* (IEEE), 1–4.

Scheutz, M., and Schermerhorn, P. (2009). "Affective goal and task selection for social robots," in *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, ed. J. Vallverdú (IGI Global) 74.

Singer, P. (2015). *The Logic of Effective Altruism*. Boston Review. Available at: bostonreview.net/forum/peter-singer-logic-effective-altruism

Sorell, T., and Draper, H. (2014). Robot carers, ethics, and older people. *Ethics Inf. Technol.* 16, 183–195. doi:10.1007/s10676-014-9344-7

Urmson, J. O. (1958). "Saints and heroes," in *Essays in Moral Philosophy*, ed. A.Melden (Seattle: University of Washington Press), 198–216.

Versenyi, L. (1974). Can robots be moral? *Ethics* 84, 248–259. doi:10.1086/291922

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.